MEMORANDUM

Re: Predicting, at posting time of a project, if a project will not get fully funded so we can intervene and help them improve the project listing.

I. Data exploration

We used two datasets provided by DonorsChoose.org: Projects and Outcomes. We indexed the datasets by their unique project IDs, and then joined them together in a new dataset. The resulting dataset (donors_choose) contained a full list of projects, broken out by descriptive variables, classified as those that received full funding and those that did not.[1]

A summary of descriptive statistics showed that 70% of the projects were chosen by donors to be fully funded. Looking at projects that were funded versus those that were not, two variables had notable differences in mean: Total Price Including Optional Support and Total Price Excluding Optional Support. Those projects that were select to be fully funded on average had a lower total price.

A correlation matrix showed that projects which were 'eligible double your impact match' were the most highly correlated with being fully funded in comparison to other project variables (correlation of 0.12).

Projects which tend to apply to DonorsChoice.org for funding follow these characteristics: urban schools; teachers with the prefix 'Mrs.'; primary focus area in Literacy & Language; resource type is supplies; categorized as highest poverty; and grade level is PreK-2. This may be useful if we want to expand our targeting of schools with different compositions, which may also attract donors of different invested interests.

II. Evaluation at the top 20%

Having completed our initial scoping of the variables, we tested various classifiers and their strength at making predictions on the likelihood of a project being funded. We included temporal validation methods, training and testing our classifiers over an expanding timeframe.

Looking at the last temporal split, where all of the data was included in the train and test sets of the classifiers, Logistic Regression returned the highest accuracy score of all of

[1] The variable we were trying to predict was fully_funded. The data was limited to a timeframe between 2011 and 2013.

the classifiers (measured by AUC-ROC of 0.557457).[2] This value, however, is very low and suggests that the classifiers were weakly discriminating.

The models with the highest precision at the top 20% were Random Forests, Boosting and Decision tree classifiers (precision ~1); the classifiers with highest recall/sensitivity at the top 20% were Random Forests, Boosting and Decision Tree (recall ~0.27). The models with the highest precision at the top 50% were Boosting, Random Forests and Decision tree classifiers (precision ~1); the classifier with highest recall/sensitivity at the top 50% was Random Forests (recall ~0.69).

Regardless of temporal splits, all classifiers tended to have low AUC-ROC scores in the range of 0.5-0.6. Precision remained relatively the same over time, with Random Forests, Boosting, and Decision Trees as the most precise at top 20% and top 50% of all the classifiers. Recall at the top 20% remained the same over time, and there were significant increases in Recall when expanded to the top 50%, which also remained consistent over time with Bagging, Random Forest, and Decision trees at the best performing.

III. Recommendations

Reviewing evaluation metrics of precision, regardless of the timeframe and regardless of the size of the population we want to intervene on, we suggest using Random Forests, Decision Trees or Boosting classifiers to predict the likelihood of being fully funded. With increased population size for intervention, recall increases for all of these classifiers.

[2] This was closely followed by Naïve Bayes (AUC-ROC = 0.557008) and Bagging (AUC-ROC = 0.555450)

First Split: Classifier with highest…		
Precision at 20%	Boosting	0.96
	Random Forests	0.96
	Decision Tree	0.96
Recall at 20%	Boosting	0.28
	Random Forests	0.28
	Decision Tree	0.28
Precision at 50%	Boosting	0.98
	Random Forests	0.98
	Decision Tree	0.98
Recall at 50%	Boosting	0.7
	Random Forests	0.7
	Decision Tree	0.7
Second Split: Classifier with highest…		
Precision at 20%	Random Forests	1
	Boosting	0.96
	Decision Tree	0.96

Recall at 20%	Random Forests	0.27
	Boosting	0.26
	Decision Tree	0.26
Precision at 50%	Boosting	0.96
	Decision Tree	0.96
Recall at 50%	Boosting	0.66
	Decision Tree	0.66
Third Split: Classifier with highest…		
Precision at 20%	Decision Tree	0.96
	Boosting	0.96
Recall at 20%	Decision Tree	0.28
	Boosting	0.28
Precision at 50%	Decision Tree	0.98
	Boosting	0.98
Recall at 50%	Decision Tree	0.71
	Boosting	0.71
Fourth Split: Classifier with highest…		
Precision at 20%	Random Forest	1
	Decision Tree	0.96
	Boosting	0.96

Recall at 20%	Random Forest	0.28
	Decision Tree	0.27
	Boosting	0.27
Precision at 50%	Random Forest	1
	Decision Tree	0.98
	Boosting	0.98
Recall at 50%	Random Forest	0.69
	Decision Tree	0.68
	Boosting	0.68